MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2017/2018

TMM 3341– DATA MINING AND MACHINE LEARNING SYSTEMS

(All sections / Groups)

28 OCTOBER 2017
2.30pm – 4.30pm
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 3 pages with 5 Questions only.
2. Answer all **FIVE** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

a. What is data mining? Is it a simple transformation of technology developed from statistics, machine learning and databases? Explain why or why not. (4 marks)

b. Define **THREE (3)** characteristic of data quality and briefly describe these concepts. (3 marks)

c. What are the differences between stepwise forward selection and stepwise backward elimination for feature (variable) selection? Answer with examples. (3 marks)

QUESTION 2

a. Suppose that the data for analysis include the attribute price (RM). The price values for the data tuples are:

33, 14, 16, 18, 20, 21, 52, 22, 70, 22, 25, 25, 40, 30, 33, 25, 33, 16, 35, 21, 35, 25, 13, 45,

i. Use smoothing by bin means to smooth the above data, using bins of depth 3. Elaborate your steps. (4 marks)

ii. Transform the price value 35 onto the range [0.0, 1.0] using min-max normalization. (2 marks)

b. What is the main difference between nominal and ordinal attributes? Elaborate how this differences affect similarity assessment. (4 marks)

QUESTION 3

a. Construct a fully connected feed forward network that accepts three inputs values, has one hidden layer of five nodes, and an output layer containing four nodes. (4 marks)

b. Given a breast tumor identification system, formulate the Task to be learned (T), performance measurement (P) and the experience based on (E). (3 marks)

c. Because many data mining algorithms cannot handle missing values, sometimes all observations (rows) that contain missing values are removed by researches before analysis. Discuss **TWO (2)** potential disadvantages of this procedure. (3 marks)

Continued .....

QUESTION 4

a. Given two objects represented by the tuples (20, 3, 36, 10) and (18, 2, 30, 8):

i. Measure the Euclidean distance between the two objects.

ii. Measure the Minkowski distance between the two objects, using p = 3. (3 marks)

b. The following contingency table summarizes supermarket transaction data.

	chickens	$\overline{\text{chickens}}$	Σ_{row}
nuggets	4000	1000	5000
$\overline{nuggets}$	2000	3000	5000
Σ_{col}	6000	4000	10000

i. Suppose that the association rule "chickens ⇒ nuggets" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, determine if this rule has strong association? (3 marks)

ii. Based on the given data, evaluate whether the purchase of chickens is independent of the purchase of nuggets. If not, what kind of correlation relationship exist between the two? (4 marks)

QUESTION 5

a. A test for a disease gives a correct positive result with a probability of 0.95 when the disease is present, but gives an incorrect positive result (false positive) with a probability of 0.15 when the disease is not present.

If 5% of the population has the disease, and Jean tests positive to the test, evaluate what is the probability Jean really has the disease? (4 marks)

b. Classifier A has 90% accuracy on the training set and 75% accuracy on the test set. Classifier B has 78% accuracy on both the training and test sets. Therefore, we can conclude that classifier A is better than classifier B (because it has better mean accuracy). True or False. Explain. (3 marks)

c. Suppose that there are a total of 50 data mining-related documents in a library of 200 documents. Suppose that a search engine retrieves 10 documents after a user enters 'data mining' as a query, of which 5 are data mining-related documents. Measure the precision and recall. (3 marks)

End of Page